Machines and Algorithms
http://www.knovell.org/mna

*Research Article,*

# Automated Detection of Diabetic Retinopathy Using Vision Transformers: A Deep Learning Approach

**Tauseef Ahmad**[1, *]

[1]IAG Software House, Multan, 66000, Pakistan

[*]Corresponding Author: Tauseef Ahmad. Email: tauseephahmad1@gmail.com

Received: 02 January 2025; Revised: 25 February 2026; Accepted: 03 March 2026; Published: 16 March 2026

AID: 005-01-000064

**Abstract:** Diabetic Retinopathy (DR) is a prevalently occurring impediment of diabetes mellitus. It is also considered to be one of the leading causes of visual impairment in the world. The importance of timely intervention and treatment is vital in finding the stages of DR and classifying them accordingly in the initial stages of detection. This study enhances a deep learning resolution on Vision Transformers (ViT) in image automated detection and grading of DR on retinal fundi images. The proposed model processes high-resolution retinal images by dividing them into patches before applying the self-attention mechanisms to identify more intricate features that indicate the extent of the DR severity by relying on the APTOS 2019 Blindness Detection dataset. The measures of the model performance are accuracy, precision, recall, and F1-score which demonstrate the efficiency of the model in classifying the different stages of DR. The results indicate the potential of ViT-based architectures in enhancing the procedure of screening offered by DR, which is a promising method in ophthalmic diagnosis.

**Keywords:** Deep Learning; Vision Transformer; Diabetic Retinopathy Diagnosis; Fundus Imaging; Automated Screening; Retinal Image Classification;

## 1. Introduction

DR is a microvascular disorder of long-term hyperglycemia among diabetic patients, which results in destroying the blood vessels in the retina. As the rate of diabetes has been rising among the global population, DR has been one of the health challenges that have gained a lot of seriousness due to the condition being among the foremost causative agents of avertible blindness in the working-age adults. The conventional approach to diagnosis of DR comprises the ophthalmologists either physically examining the retinal fundus images and this is time consuming and can be affected by inter-observer variability. The image analysis of medical images has been revolutionized by deep learning and Convolutional Neural Networks (CNNs) have been widely used in the detection of DR. However, CNNs are usually large in size when dealing with datasets and may experience the challenge of deriving global contextual data in images. [1], [2]

Vision Transformers have recently also been used to analyze images based on the success of transformers in natural language processing. ViTs break down images into patches and operate it using self-attention mechanisms, which enables the model to bring out local and global features effectively. This architecture has been proved in a number of computer vision applications and one of them is medical



https://doi.org/10.66108/mna.v5i01.113

imaging [3]. The current paper talks about the possibility of identifying and grading DR with the help of ViT with the APTOS 2019 [4]. We also aim at seeking the viability of ViT in the appropriate categorization of the DR phases, therefore, assisting to develop automated, reliable, and effective DR screening tools.

This paper addresses the problem of effective and automated stage of diabetic retinopathy diagnosis based on retinal fundus, especially when the medical information that can be used to label the retinal fundus stages is limited. In order to address it, a framework of a Vision Transformer is suggested, using transfer learning and patch-based image representation to extract both local and global features. The major contributions of this work are: (1) a ViT-based multi-class DR classification framework is implemented, (2) the framework is tested and evaluated on the APTOS 2019 dataset, and (3) the performance of the models is analyzed with the help of traditional classification measures and attention-based interpretability.

***Objectives***

To deal with the research problem, this research is informed by the following objectives:

- To train a Vision Transformer-based model to be used in multi-classifying diabetic retinopathy.
- To test the model on the APTOS 2019 data with the conventional performance measures.
- To compare the efficiency of ViT on international and national retinal features.
- To estimate model interpretability with the help of visualizing attention maps.

## 2. Related Work

### *2.1. Traditional Approaches to Diabetic Retinopathy Detection*

The retinal fundus images of patients with diabetes mellitus have been a conventional way of diagnosis of diabetic retinopathy, which is a microvascular complication. It is an established process, however, takes a long time, is labor-intensive and can be subject to inter-observer variance especially in mass screening or where specialist care is less accessible. The early approaches of calculation attempted to automate the DR identification procedure by extracting hand-crafted features, namely blood vessel morphology, microaneurysms, exudates and hemorrhages and apply typical classifiers, including Support Vector Machines (SVM) or Decision Trees or k-Nearest Neighbors (KNN) [5]. These methods were fairly effective; however, they were also highly susceptible to noise, changes in image quality and in most instances lacked the durability needed to be implemented to the clinical setting.

### *2.2. Convolutional Neural Networks in DR Detection*

The deep learning and specifically CNNs revolutionized the field of medical image analysis with the potential of self-learned end-to-end features as a direct product of raw pixel data. An additional investigation by [6] developed a profound CNN-based framework, which is prepared on over 120,000 retinal photographs and applied it to detect referable DR at the level of board-certified ophthalmologists. Similarly, a five-convolutional and three-fully connected CNN model was used to stage DR on the Kaggle eyePACS database and has an overall average classification rate of approximately 75 percent. [7]

Subsequent literature employed more sophisticated and improved forms of CNN such as ResNet, DenseNet, InceptionNet and EfficientNet which were discovered to be stronger and more precise on other datasets. The process of generalization was also improved by data augmentation techniques and ensembling. CNN based models are however also limited by nature of their local receptive fields whereby they can only be restricted to establish global structural relationships when provided with retinal images. Additionally, they are inclined to operate with the premise of huge volumes of labeled data that are not necessarily available in the field of medicine.

### *2.3. Vision Transformers in Computer Vision*

Transformers, which were initially designed to carry out sequential modeling procedures in natural language processing [8], have now been used in computer vision. The proposed architecture of the ViT by

[9] is not based on the traditional CNNs as the image is considered as sequential collection of fixed-size patches and the self-attention models are used in order to capture the long-range dependencies. The plan allows ViTs to recognize local and global contextual data without convolutional functions.

Despite the heavy requirements of ViTs on large image datasets (e.g. ImageNet-21k or JFT-300M) to achieve the best results, transfer learning enabled the practical use of ViTs in data-limited scenarios. ViTs are also demonstrated to outperform CNNs in a number of image classification tasks when it is needed to capture high-level structural patterns and global dependencies.

### *2.4. Vision Transformers in Medical Imaging*

Transformer based architecture to medical imaging is an emerging but rapidly evolving study. Several articles have reported on hybrid architectures that combine the local feature extraction capability of CNNs with the global reasoning capability of transformers. Indicatively, author in [10] have introduced a medical image segmentation hybrid architecture, TransUNet, which replaces ViT blocks with the U-Net architecture. Author of [11] applied a model based on transformers to divide the retina vessels into the ophthalmology domain and demonstrated higher rates of sensitivity and specificity.

Recent studies by author of [12] implemented Swin Transformer, a hierarchical vision transformer model, on this task of DR grading and stated that it performed better than the typical CNNs, which are based on classification accuracy. However, in these studies they are typically intricate hybrid architectures, or they focus on segmentation rather than classification activities. Moreover, relevance of transformer-based models to medical cases remains a research part, particularly, correspondence of the attention map with the clinically relevant features of the retina.

According to author of [6], the paper created a deep learning framework using convolutional neural networks (CNNs) to identify diabetic retinopathy in retinal fundus images. The model was trained on huge-scale datasets like EyePACS and tested on Messidor-2 with high sensitivity and specificity being comparable to those of ophthalmologists. This paper has shown that deep learning has proven effective in automated DR screening and it heavily depends on large annotated datasets and CNN-based architectures.

Author of [7], suggested a CNN-based diabetic retinopathy classification model on the basis of retinal images provided by the Kaggle dataset. Their result produced a rate of about 75% and the viability of deep learning in the staging of DR. Nevertheless, the model was poor at generalization and experimented with complex feature extraction because of its use of convolutional architectures.

Author of [13], examined ViT in diabetic retinopathy grading and showed that transformer-based models are capable of extracting global contextual information in the retinal image. Their publication emphasized the superiority of self-attention mechanisms to CNNs in long-range dependency modelling but the method used large amounts of computational resources and large-scale training data.

Author of [14] suggested FunSwin, a Swin Transformer-based architecture to evaluate diabetic retinopathy grades and macular edema. This model produced a precision of around 90.6 and was also superior to normal CNNs as it made use of hierarchical attention systems. The architecture is however rather complex and computationally intensive.

More recent models like STMF-DRNet presented in [15], provide a multi-branch Swin Transformer V2 classification framework of fine-grained diabetic retinopathy. The model incorporates global and local features extracting with the help of attention and shows higher robustness on various datasets, such as APTOS and DDR. Although the performance of the model is high, the complexity of the model and its cost of computation are main issues that need to be addressed to make it work.

## 3. Research Gap and Motivation

Nevertheless, although more and more studies focus on the application of vision transformers in healthcare, there is not much literature regarding the application of pure ViT architectures to the problem of the classification of diabetic retinopathy. Most of the literature that is available uses CNN-based designs or hybrid designs, which consist of transformer components. Also, the research lacking a critical evaluation

of ViTs on the APTOS 2019 Blindness Detection dataset which brings additional challenges in the form of the uneven quality of the image, lighting-induced artifacts, and imbalance in classes is lacking.

The gaps in this paper are addressed with the analysis of a ViT-based automated DR recognizer and grader. Our approach is a transfer learning approach, whereby pre-trained ViT weights are used and the model is optimized on APTOS-based high-resolution fundus images. Our target is to consider the performance of vision transformers on local and global pathology of DR, and therefore offer a scalable and robust solution that is comparable to conventional CNN-based algorithms.

As opposed to Swin Transformer-based methods, which use hierarchical architecture, this paper will concentrate on a pure Vision Transformer (ViT) model. This enables a much simpler architecture, yet is able to obtain competitive performance without such complex hybrid or hierarchical designs.

## 4. Methodology

This part gives the description of the dataset, data preprocessing strategies, model architecture, training pipeline, and evaluation metrics that were adopted to build and evaluate a Vision Transformer model to find Diabetic Retinopathy in an automated way. All the implementation was carried out in PyTorch in Kaggle where training and evaluation occurred via graphical processor acceleration. Figure 1 below depicts proposed methodology diagram.

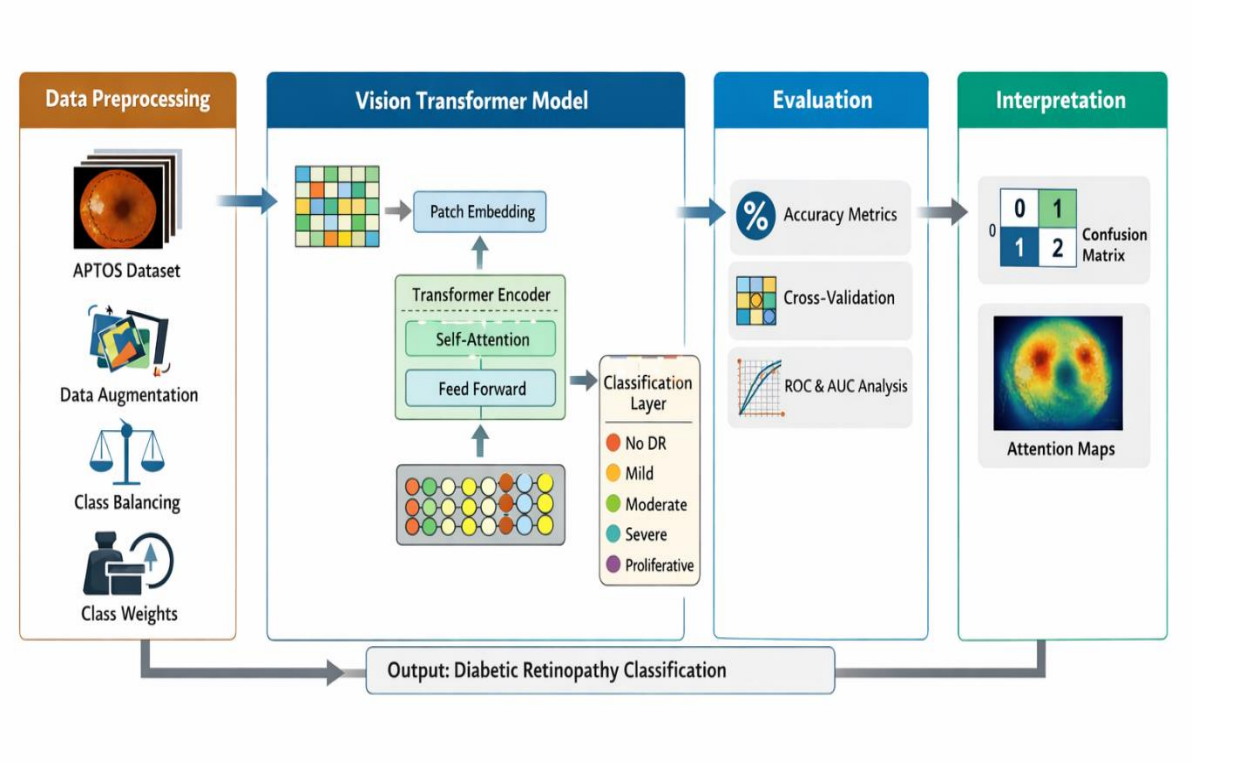


**Figure 1:** Proposed Methodology Diagram

### *4.1. Dataset Description*

The dataset to conduct this research is APTOS 2019 Blindness Detection dataset available currently on Kaggle as one of the competitions hosted by Asia Pacific Tele-Ophthalmology Society (APTOS). It is composed of 3,662 retinal fundus images of color, the images are marked according to five stages DR. [16] Figure 2 below depict two sample images of input dataset.

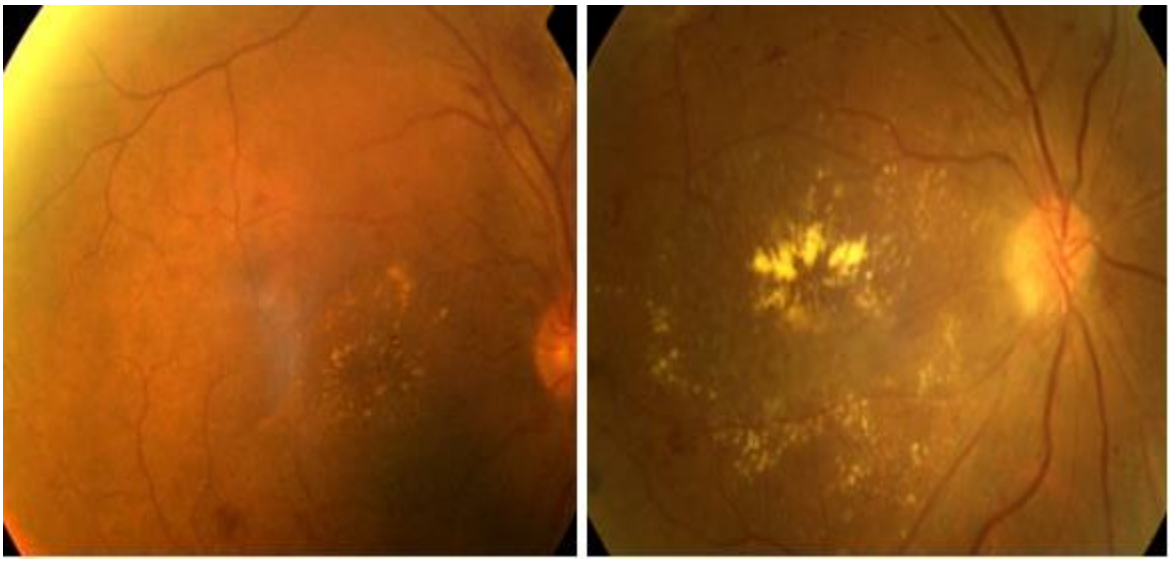

**Figure 2:** Sample APTOS dataset images

**Classification system: 0**: No DR, **1**: Mild**, 2**: Moderate, **3**: Severe**, 4**: Proliferative DR

Each image is linked to a seamlessly anonymized patient and the image is in PNG format and various resolutions, which in most cases are larger than 3000x 2000 pixels. The frequency distribution of the label is skewed and high frequency of class 0 (No DR), which has been considered in the model checking.

Stratified sampling was done with the parity of the distribution of classes across the dataset to divide it into training (80) and validation (20). The mappings between the label and the image were done using the data in the train.csv file.

### *4.2. Data Preprocessing and Augmentation*

The preprocessing was significant because the resolutions and quality of the images change significantly, which would result in the successful model convergence. It has been done through the following steps:

- **Image Resizing**: Scale: 224 by 224 pixels of the images was done by bilinear interpolation. This dimension corresponds to the input requirements of the ViT model (vit_base_patch16_224). [17]
- **Color Space Conversion**: The picture was transformed to RGB using the PIL.Image library in such a way that all the channels were fed the same way. [18]
- **Normalization**: Rates of tensors are normalized with a mean and standard deviation value of [0.5 0.5 0.5] and the values of pixels are normalized to the interval of -1:1.
- **Transformations**: To ensure that the above steps were automatic, a torch vision. transforms. Compose pipeline was used to apply all of them to both training and validation images.

Stable training is possible because of the preprocessing pipeline, and the artifacts that occur because of domain specificity become minimized, which may negatively affect model generalization. Although the data augmentation was added to the minimum, in the future, this pipeline can be extended to include random rotation, flipping, and contrast which are other data augmentation methods.

### *4.3. Custom Dataset Loader*

PyTorch DRDataset is a subclass that has been developed to facilitate the loading of the data efficiently. The class:

- Loads the name of the image and the name in the CSV annotation.

- Training is done through reading and processing each image on-the-fly.
- Gives a pair of an image and a label (image-tensor, label) of each indexed sample.
- PyTorch DataLoader perfectly matches with this class, and it shuffles, mini-batches, etc.
- Integrates with DataLoader for batching and shuffling.

### *4.4. Vision Transformer Architecture*

The model that will be involved in processing images is the ViT model, which was modified by author of [19] and is found in the timm (PyTorch Image Models) library [20]. The architecture utilized is vit_base_patch16_224 and the most critical feature of the architecture is the following:

- **Patch Embedding**: The images are segmented into 16x16 patches of all the images and the output size is 196 tokens per image.
- Transformer Encoder: Patch embeddings are fed through 12 transformer blocks that consist of multi-head self-attention, layer normalization, and MLP layers.
- **Positional Encoding**: Learner positional encodings of spatial information are stored between patches of images.
- **Classification Head**: The final CLS token is transmitted by a fully connected head, which has been programmed to furnish 5 class outputs where there is a case of DR classification.

ImageNet pre-trained weights also initiated the model and enabled transfer learning to accelerate convergence and yield improved results on medical imaging data.

### *4.5. Training Procedure*

Training of the model was done with Adam optimizer and learning rate of 1e-4. Since the objective term is the categorical cross-entropy loss, the categorical cross-entropy loss (nn.CrossEntropyLoss) was selected (it is suitable with a multi-class classification problem).

The training was conducted in 10 epochs and the performance was checked at the end of each epoch. The training loop included:

- ViT model forward propagation.
- Calculation of backward-propagation and losses.
- Optimizer Gradient updates.
- Accuracy tracking per epoch

Values of loss, classification accuracy would be saved in each of all the epochs to determine the convergence pattern.

### *4.6. Evaluation and Validation Strategy*

Validation set was used after every epoch to check the generalization capacity of the model. The evaluation was composed of:

- **Accuracy**: Fraction of the correctly labeled labels.
- **Precision, Recall, and F1-score**: The scores are computed on a per-class basis, using Scikit-learn classification report.
- **Confusion Matrix**: Visualized with the assistance of the Seaborn to view the performance of the classes and how they are misclassified.

These will provide an overall performance and class specific information that are crucial because there are clinical ramifications of under or over-anticipating some steps of the DR.

### *4.7. Attention Map Visualization*

The inner functioning of ViT model was analyzed with the help of attention map visualization. The attention weights were to be registered on the attention layers by registering forward hooks to the attention

layers during inference. A specific attention head and a layer had to be selected in order to see the distribution of self-attention patches to the image.

The resulting attention heatmaps indicate the parts of the retina that had the strongest influence in predicting the model, which is any form of explainability that is the most important in the clinical practice. These images are incorporated to ascertain that the model is centered at medically significant regions e.g. hemorrhages, microaneurysms or exudates.

### *4.8. Tools and Environment*

All of them were run on Kaggle Notebooks, where each acceleration was free on NVIDIA Tesla P100. The stack of implementation was:

- Python 3.10
- PyTorch 2.0+
- timm (Hugging Face) for ViT architecture
- Seaborn & Matplotlib for data and attention visualization
- Scikit-learn for evaluation metrics

This architecture offered a scalable and repeatable experimental environment of deep learning.

## 5. Result

The ViT model was experimented on the APTOS 2019 dataset on different performance metrics, including training loss, accuracy, precision, recall, F 1score, and confusion matrix analysis.

### *5.1. Training Progress*

The model has been trained in 10 epochs with a batch size of 16 i.e., 16 images are processed at a single time by the model. For learning rate optimization, I have exploited Adam optimizer, while the learning rate is set to 1e-4. By the analysis of results it could be clearly noticed that the training precision gradually increased as the number of epochs was increased and it was an indicator of successful model learning. By these results we can say that the model is converging appropriately.

Table 1 below summarizes the accuracy findings during model training. At the end of first epoch training accuracy was observed to be 68.57%, which gradually increase and reach 74.3% at 2$^{nd}$ epoch, 76.3 at 3$^{rd}$ epoch. The final training accuracy at 10$^{th}$ epoch is 86.16, while the final training loss has dropped down from 0.88 to 0.39 at final training epoch. This accuracy ay validation data after the training of model, amounted to 91.13 percent that is, there was a good generalization and the model has not overfitted the training data.

**Table 1:** Training Loss and Accuracy per Epoch

| Epoch | Loss | Accuracy (%) |
|---|---|---|
| **1** | 0.8755 | 68.57 |
| **2** | 0.7343 | 74.30 |
| **3** | 0.6460 | 76.32 |
| **4** | 0.6116 | 77.99 |
| **5** | 0.5674 | 79.71 |
| **6** | 0.5396 | 80.80 |
| **7** | 0.5122 | 81.81 |
| **8** | 0.4832 | 82.93 |
| **9** | 0.4232 | 85.20 |
| **10** | 0.3911 | **86.16** |

### *5.2. Classification Performance*

These were precision, recall and F1-score which detected how the model performed on the validation set with regards to classification. These measurements were determined in consideration of five categories of DR, that is, No DR, Mild, Moderate, Severe, and Proliferative DR. The detailed classification report is given in table 2.

**Table 2:** Classification Report on Validation dataset

| Class | Recall | Precision | F1-score | Support |
|---|---|---|---|---|
| **No DR** | 0.99 | 0.99 | 0.99 | 289 |
| **Moderate** | 0.89 | 0.84 | 0.86 | 160 |
| **Mild** | 0.80 | 0.81 | 0.80 | 59 |
| **Proliferative DR** | 0.85 | 0.89 | 0.87 | 47 |
| **Severe** | 0.58 | 0.75 | 0.65 | 31 |
| **Overall (Accuracy)** | | | **0.91** | **586** |

The model was also very precise and precise in the case of the No DR and the Moderate classes. However, the category of Severe had slightly lower recall, and this means that it was marginally hard to draw the line between it and other similar stages, including the ones of the classes of the Moderate and the Proliferative DR.

The improved performance of the higher-than-average performance of the No DR and the Moderate classes can be explained by the fact that they have a greater representation in the data and thus allow the model to learn more robust features. On the contrary, the poor recall of the “Severe” group indicates imbalance of classes and subtle differences between the features of the adjoining stages, which makes the classification more difficult.

### *5.3. Confusion Matrix Analysis*

Another detail of the classification behavior of the model is provided in the confusion matrix (Figure 3). A majority of the misclassifications were of two classes that were closely similar e.g., Mild and Moderate or Moderate and Severe, which is understood to be a challenge in the DR detection as there are very slight dissimilarities in retinal pathology.

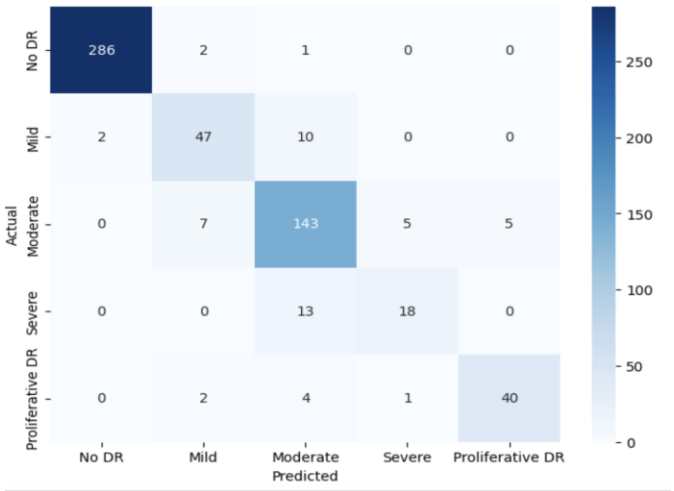


**Figure 3:** Confusion matrix showing classification performance across DR stages

As it is mentioned in the model, images in the no DR category were correctly identified as 286 of the 289 and those in the moderate category were correctly identified as 143 of the 160, and the result is high in the most used and relatively popular classes.

### *5.4. Comparison with Existing Studies*

Table 3 below illustrate the comparison of proposed model with existing studies.

**Table 3:** Comparison of Proposed Model with Existing Studies

| Study | Model | Dataset | Performance | Approach |
|---|---|---|---|---|
| **Gulshan et al. (2016)** | CNN (Inception-v3) | EyePACS + Messidor-2 | AUC: 97% | Deep CNN |
| **Pratt et al. (2016)** | CNN | Kaggle DR dataset | Accuracy: ~75% | CNN |
| **Wu et al. (2021)** | Vision Transformer | Fundus dataset | Improved classification | ViT |
| **Yao et al. (2022)** | Swin Transformer (FunSwin) | DR datasets | Accuracy: 90.6% | Transformer |
| **STMF-DRNet (2024)** | Swin Transformer V2 | APTOS + DDR | Accuracy: ~86% | Hybrid Transformer |
| **Proposed Work** | ViT | APTOS 2019 | Accuracy: 91.13% | Pure ViT |

## 6. Discussion

The article examines the application of ViTs to stage diabetic retinopathy in APTOS 2019 in Blindness Detection dataset. Our model was approximately 90 percent valid which demonstrates how ViTs could depict long-range dependencies in retinal images. ViTs have a self-attention mechanism, which enables the model to focus on the useful portions of the retina and the proper DR classification is necessary to take into account.

However, there were several problems encountered during the research. The distribution of the class in the dataset APTOS is not even as the majority of pictures are included in the category of No DR. In case of an unbalanced approach, prejudiced forecasts on behalf of the majority class may be present. To overcome this, weighting and oversampling of minority classes in classes was considered.

In addition, even though the ViTs are superior to the conventional Convolutional Neural Networks in terms of global context representation, they are computationally costly and require massive amounts of data to be trained. To address the issue of data scarcity in this paper, the challenge was addressed by transfer learning with the help of a pre-trained ViT model. The fact that the model can be extrapolated to the unobservable data is still a problem and hence further testing on the external data is still essential.

The other emphasis was that ViT model should be interpretable. Even though the focus can be provided through the attention maps, which can provide the notion of the areas, which the model is focused on, the decision-making process cannot be fully comprehended. Future research should aim to find the means of enhancing interpretability of ViTs to help implement them in clinical practice.

## 7. Conclusion

The aim of the present study was to create a system to classify diabetic retinopathy with the help of a Vision Transformer-based system. Through transfer learning and patch-based feature extraction, both the local and global retinal features were well captured by the model. The experimental results on the APTOS 2019 dataset showed that there was a high degree of classification in that it achieved a high level of accuracy and was likely to predict DR stage-wise.

These findings highlight the necessity to address the problem of the data imbalance, further the model interpretability, and test the models on extensive data to be robust. As the sphere changes, the introduction of ViTs into clinical practice would make the process of DR screening to be identified earlier and result in better patient outcomes.

## 8. Future Work

Future research will concentrate on the enhancement of the data balance with the help of sophisticated augmentation and synthetic data creation methods. Also, computational efficiency will be optimized in Vision Transformer models, which will allow applying them to resource-limited settings. More studies will be conducted to consider federated learning deviations to guarantee data privacy, and more powerful interpretability deviations to promote clinical trust. Lastly, it will be necessary to conduct a large-scale clinical validation to determine the real-world applicability and strength.

**Funding Statement:** Author has not received any funding.

**Conflicts of Interest:** NIL.

**Data Availability:** The dataset used in this study is publicly available on Kaggle.

## References

[1] Gettinger, K., Lee, D., Tomita, Y., Negishi, K., & Kurihara, T. (2025). Diabetic Retinopathy, a Comprehensive Overview on Pathophysiology and Relevant Experimental Models. International Journal of Molecular Sciences, 26(20), 9882. https://doi.org/10.3390/ijms26209882

[2] Fong, D. S., Aiello, L. P., Ferris, F. L., & Klein, R. (2004). Diabetic Retinopathy. Diabetes Care, 27(10), 2540–2553. https://doi.org/10.2337/diacare.27.10.2540

[3] Wang, A. (2025). Vision Transformers (ViTs): A New Era in Computer Vision – A Review. Journal of Computing and Electronic Information Management, 18(3), 23–30. https://doi.org/10.54097/41t0vx90

[4] Awasthi, V., Awasthi, N., Kumar, H., Singh, S., Singh, P. P., Dixit, P., & Agarwal, R. (2024). ViT-HHO: Optimized vision transformer for diabetic retinopathy detection using Harris Hawk optimization. MethodsX, 13, 103018. https://doi.org/10.1016/j.mex.2024.103018

[5] Mhasawade, A., Rawal, G., Roje, P., Raut, R., & Devkar, A. (2023). Comparative study of SVM, KNN and Decision Tree for Diabetic Retinopathy Detection. 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 166–170. https://doi.org/10.1109/cises58720.2023.10183456

[6] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402. https://doi.org/10.1001/jama.2016.17216

[7] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional Neural Networks for Diabetic Retinopathy. Procedia Computer Science, 90, 200–205. https://doi.org/10.1016/j.procs.2016.07.014

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N.Gomez, A., Kaiser, L., & Polosukhin, I. (2025). Attention Is All You Need. https://doi.org/10.65215/r5bs2d54

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[10] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[11] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[12] Mallampalli, P., & R, M. (2025). Diabetic Retinopathy Classification Using Swin Transformer: A Vision Transformer-Based Approach. 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), 1181–1188. https://doi.org/10.1109/icici65870.2025.11069800
[13] Wu, J., Hu, R., Xiao, Z., Chen, J., & Liu, J. (2021). Vision Transformer-based recognition of diabetic retinopathy grade. Medical Physics, 48(12), 7850–7863. Portico. https://doi.org/10.1002/mp.15312
[14] Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., & Wang, Z. (2022). FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. Frontiers in Physiology, 13. https://doi.org/10.3389/fphys.2022.961386
[15] Liu, Y., Yao, D., Ma, Y., Wang, H., Wang, J., Bai, X., Zeng, G., & Liu, Y. (2025). STMF-DRNet: A multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2. Biomedical Signal Processing and Control, 103, 107352. https://doi.org/10.1016/j.bspc.2024.107352
[16] Karthik, Maggie, & Dane, S. (2019). APTOS 2019 Blindness Detection [Data set]. Kaggle. https://www.kaggle.com/competitions/aptos2019-blindness-detection
[17] Halder, A., Gharami, S., Sadhu, P., Singh, P. K., Woźniak, M., & Ijaz, M. F. (2024). Implementing vision transformer for classifying 2D biomedical images. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-63094-9
[18] Mohaideen Abdul Kadhar, K., & Anand, G. (2024). Python Libraries for Image Processing. Industrial Vision Systems with Raspberry Pi, 33–72. https://doi.org/10.1007/979-8-8688-0097-9_3
[19] Dosovitskiy, A., & Brox, T. (2016). Inverting Visual Representations with Convolutional Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4829–4837. https://doi.org/10.1109/cvpr.2016.522
[20] Mishra, P. (2019). CNN and RNN Using PyTorch. PyTorch Recipes, 49–109. https://doi.org/10.1007/978-1-4842-4258-2_3